

Faculty of Engineering and Information Technology
University of Technology Sydney

Data Mining for High Performance Compression of Genomic Reads and Sequences

A thesis submitted in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

by

Yuansheng Liu

September 2019

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, Yuansheng LIU declare that this thesis, is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the Faculty of Engineering and Information Technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise reference or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Production Note:

Signature: Signature removed prior to publication.

Date: 1/10/2019

Acknowledgments

First and foremost, I offer my deepest gratitude to my supervisor, Professor Jinyan Li, for his extensive support and guidance during last three years of my PhD studies. With his help and strong support, I had a chance of winning scholarships. Without his patience, I never would have been uncovered the fascinating areas of bioinformatics. His guidance including scientific writing and research ideas made all of this work possible.

I am very thankful to Dr. Hui Peng, who is our team member and is also my roommate at Xiangtan University. I could not have the opportunity to study at UTS without his strong recommendation. Dr. Peng has vast knowledge in the area of bioinformatics and also gives many insightful suggestions and comments to improve the quality of our projects. I would also like to thank other team members: Chaowang Lan, Tao Tang, Xiaocai Zhang, Xuan Zhang, Yi Zheng and Zhixun Zhao, for their kind help during my studies. And the delicious food cooked by them is really evocative. The fantastic activities with them are enjoyable and relaxing under the heavy pressure of research.

This work would not be possible without the people with whom I collaborated and whose help was essential in many projects I participated in: Prof. Limsoon Wong, Prof. Zuguo Yu, Prof. Quan Zou, Prof. Xiangxiang Zeng and Dr. Yu Zhang.

Furthermore, I gratefully acknowledge all the organizations provided the scholarships including International Research Scholarships, ARC Discovery Scholarship, and the Big Data Big Impact grant.

Acknowledgments

I am deeply grateful to my wonderful parents and my dear wife for their support and love throughout all my life. I also extend my deep gratefulness to my relatives and friends in China.

As to my lovely daughter and son, my love for them is beyond words. I am very sorry for not being able to accompany in the first two years of them as I study overseas.

Yuansheng Liu

UTS, Sydney, Australia

June 2019

Contents

Certificate	i
Acknowledgment	iii
List of Figures	ix
List of Tables	xi
List of Publications	xiii
Abstract	xv
 Chapter 1 Introduction	 1
1.1 Compression of Raw Sequencing Data	3
1.2 Compression of Assembled Genomes	5
1.3 Detection of Maximal Exact Matches	6
1.4 Research Contributions	7
1.5 Thesis Organization	9
 Chapter 2 Related Work and Literature Review	 10
2.1 Short Reads Compression	10
2.2 Reference-based Genome Compression	13
2.3 Detection of Maximal Exact Matches	16
 Chapter 3 Index Suffix-prefix Overlaps by (w, k)-minimizer to Generate Long Contigs for Reads Compression	 18
3.1 Introduction	18
3.2 Materials and methods	20
3.2.1 Reads indexing and iterative contigs merging	22
3.2.2 Realignment of singleton reads	27

3.2.3	Encoding	27
3.2.4	Reads-order preserving mode	29
3.2.5	Handling paired-end reads	30
3.2.6	Other considerations	30
3.3	Results and analysis	31
3.3.1	Datasets	32
3.3.2	Compression performance	32
3.3.3	Comparison on computational resources	40
3.4	Conclusion	41

Chapter 4 High-speed and High-ratio Referential Genome

	Compression	42
4.1	Introduction	42
4.2	Materials and methods	44
4.2.1	Preprocessing steps	44
4.2.2	Advanced greedy matching in a global hash table	46
4.2.3	Post-processing	48
4.2.4	Decompression	48
4.3	Results and performance analysis	50
4.3.1	Genome data sets and their disk file size	50
4.3.2	High compression ratios on the 8 benchmark human genomes	51
4.3.3	Time complexity and speed performance	58
4.3.4	Memory usage by HiRGC	61
4.3.5	Compressing 100 genomes from the 1000 Genomes Project	61
4.3.6	Performance on compressing the genomes of some plants and microbial species	62
4.4	Conclusion	62

Chapter 5	Fast Detection of Maximal Exact Matches via Fixed Sampling of Query k-mers and Bloom Filtering of Index k-mers	63
5.1	Introduction	63
5.2	Materials and methods	64
5.2.1	bfMEM algorithm	64
5.2.2	Bloom filter and rolling hash	72
5.3	Results and analyses	74
5.3.1	Datasets	75
5.3.2	Comparison on the number of generated query k -mers .	75
5.3.3	Comparison on the number of indexed k -mers	75
5.3.4	Running time and memory usage comparison	76
5.3.5	Effect of multi-threads	81
5.4	Conclusion	82
Chapter 6	Conclusions and Future Work	83
6.1	Conclusions	83
6.2	Future Work	84
Chapter A	Supplementary files	86
Bibliography	87

List of Figures

1.1	An entry from the WGS data SRR327342.	4
3.1	An illustration of contigs merging via suffix-prefix overlaps. . .	25
4.1	Schematic diagram of our algorithm HiRGC.	45
4.2	Box plots of compressed file sizes by different methods.	56
5.1	Schematic diagram of our algorithm bfMEM.	66
5.2	Fixed sampling of k -mers.	67
5.3	An illustration of adding and querying elements to the Bloom filter.	73

List of Tables

3.1	Sizes (in KB) of the compressed files in the compression of single-end FASTQ files	33
3.2	Size (in byte) of compressed files for paired-end reads	36
3.3	Sizes (in KB) of compressed files in the reads-order preserving mode	39
4.1	Overall comparison of compressed size (MB) and relative compression gain for different algorithms under different reference genomes.	52
4.2	Detailed comparison between HiRGC and the state-of-the-art methods	53
5.1	Comparison on the numbers of query k -mers used by E-MEM, copMEM and bfMEM	77
5.2	Comparison on the numbers of indexed k -mers used by E-MEM, copMEM and bfMEM	77
5.3	Comparison of running times (in second) by different methods for $L = 40, 50, 80$ and 100	78
5.4	Comparison of memory usage (in GB) by different methods for $L = 40, 50, 80$ and 100	79
5.5	Comparison of running time (in second) and memory usage (in GB) by different methods for $L = 150, 200$ and 300	80

List of Publications

Below is the list of journal papers associated with my PhD research:

Journal Papers Published

- **Liu, Y.**, Zhang, Y., & Li, J. (2019). Fast detection of maximal exact matches via fixed sampling of query k -mers and Bloom filtering of index k -mers, *Bioinformatics*, in press: 10.1093/bioinformatics/btz273.
- **Liu, Y.**, Yu, Z., Dinger M.E., & Li, J. (2019). Index suffix-prefix overlaps by (w, k) -minimizer to generate long contigs for reads compression. *Bioinformatics*, 35(12), pp.2066–2074.
- **Liu, Y.**, Peng, H., Wong, L., & Li, J. (2017). High-speed and high-ratio referential genome compression. *Bioinformatics*, 33(21), pp.3364–3372.
- **Liu, Y.**, Lan, C., Blumenstein, M., & Li, J. (2017). Bi-level error correction for PacBio long reads. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, in press: 10.1109/TCBB.2017.2780832.

Abstract

The rapid development of next-generation sequencing (NGS) technologies has revolutionized almost all fields of genetics. However, the massive amount of genomic data produced by NGS presents great challenges to data storage, transmission and analysis. Among various NGS-related big data challenges, in this thesis, we focus on short reads data compression, assembled genome compression and maximal exact matches (MEMs) detection.

First we propose a new *de novo* compression algorithm for short reads data. The method utilizes minimizers to exploit the redundant information presented in reads. Specifically, large k -minimizers are used to group reads and (w, k) -minimizers are used to search suffix-prefix overlap similarity between two contigs. Our experiments show that the proposed method achieves better compression ratio than the existing methods.

Furthermore, we present a high-performance reference-based genome compression algorithm. It is based on a 2-bit encoding scheme and an advanced greedy-matching search on a global hash table. The compression ratio of our method is at least 1.9 times better than the best competing algorithm on its best case, and our compression speed is also at least 2.9 times faster.

Finally we introduce a method to detect all MEMs from pairs of large genomes. The method conducts a fixed k -mer sampling on the query sequence and the index k -mers are filtered from the reference sequence via a Bloom filter. Experiments on large genomes demonstrate that our method is at least 1.8 times faster than the best of the existing algorithms.

Abstract

Overall, this thesis work has developed efficient algorithms for pattern discovery from and for data compression of genomic sequences of big size.